Linux in Production

Ron Brightwell

Sandia National Laboratories

Scalable Computing Systems

Outline

- Sandia system software research
- Target system architecture
- Cplant™ architecture
- Linux results and observations
- Issues
- Summary
- Future

Sandia System Software Research

- Intel Paragon
  - 1,890 compute nodes
  - 3,680 i860 cpu's
  - 143/184 GFLOPS
  - 175 MB/sec network
- SUNMOS lightweight kernel
  - High performance compute node OS for distributed memory MPP's
  - Deliver as much performance as possible to apps
  - Small footprint
  - Started in January 1991 on the nCUBE-2 to explore new message passing schemes and high-performance I/O
  - Ported to Intel Paragon in Spring of 1993

- Intel TeraFLOPS
  - 4,576 compute nodes
  - 9,472 Pentium II cpu's
  - 2.38/3.21 TFLOPS
  - 400 MB/sec network
- Puma lightweight kernel
  - Multiprocess support
  - Modularized (kernel, PCT)
  - Developed on nCUBE-2 in 1993
  - Ported to Intel Paragon in 1995
  - Ported to Intel TFLOPS in 1996 (Cougar)
  - Portals 1.0
    - User/Kernel managed buffers
  - Portals 2.0
    - Avoid buffering and mem copies

Target Architecture

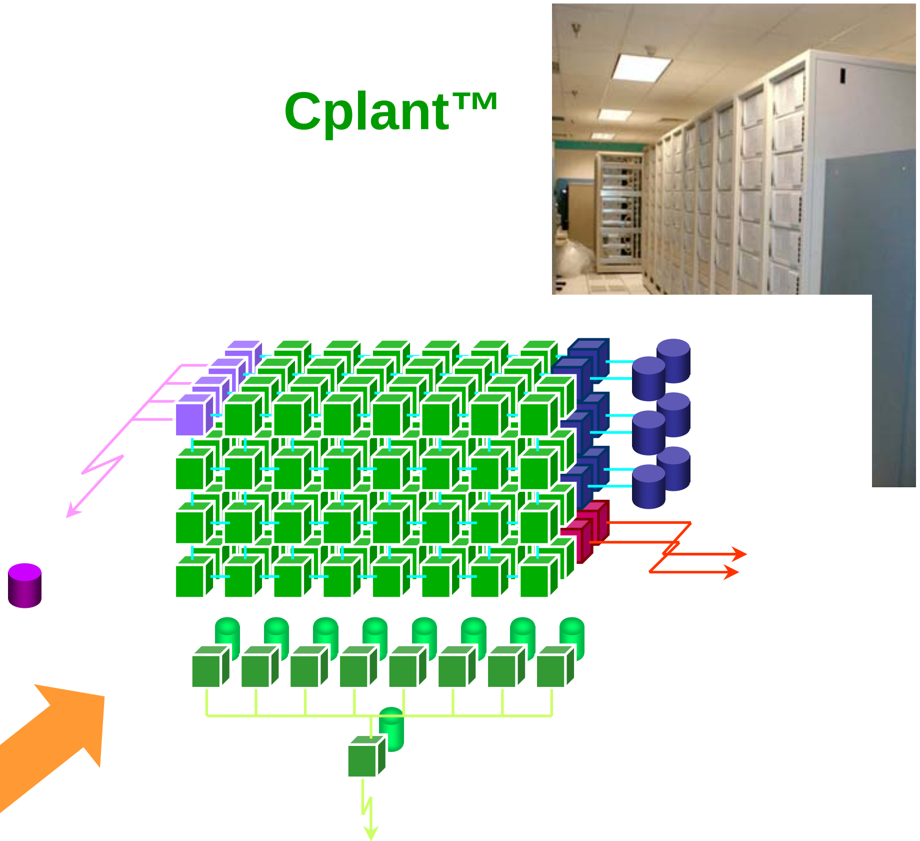
- Distributed memory message passing
- Partition model of resources
  - Compute nodes
    - Small number of CPUs
    - Diskless
    - High performance network
      - Peak processor speed in MHz is near peak network bandwidth in MB/s
  - Service nodes
  - Disk I/O nodes
  - Network I/O nodes

Cplant™ Architecture

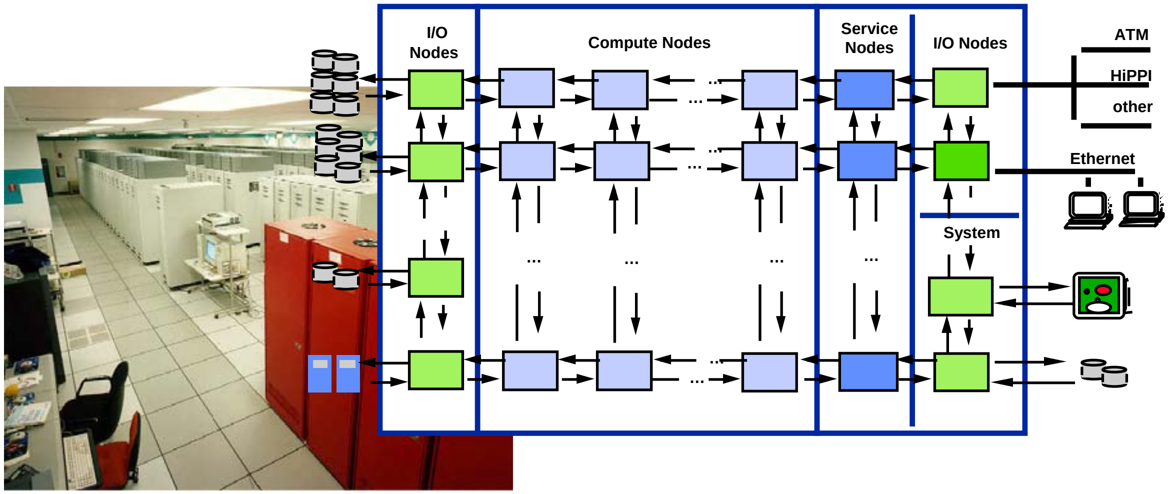
MPP “look and feel”

- Distributed systems and services architecture
- Scalable to 10,000 nodes
- Embedded RAS features
- Preserve application code base

Cplant™

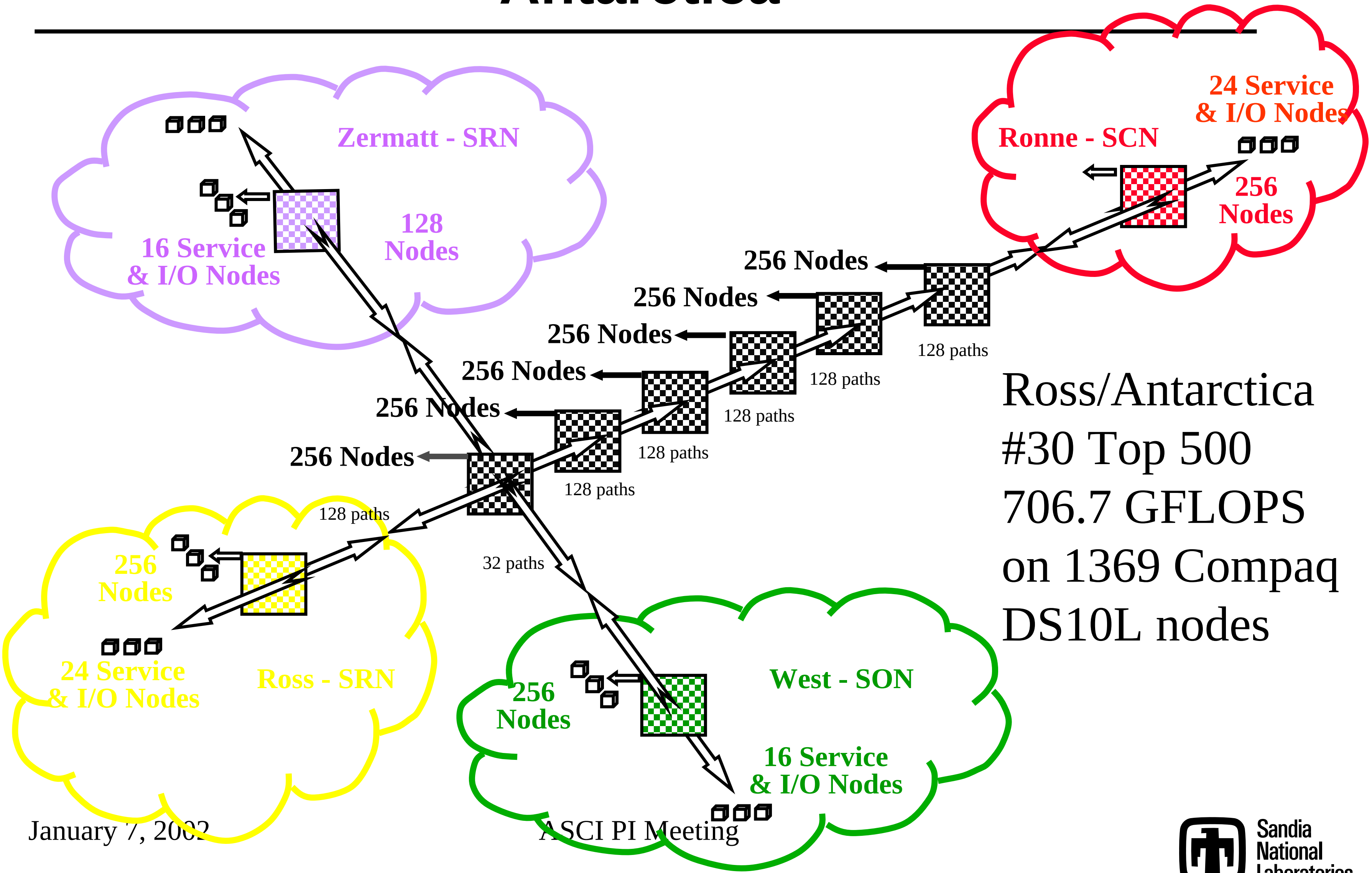


ASCI Red



Extends ASCI Red advantages

Antarctica



Ross/Antarctica
#30 Top 500
706.7 GFLOPS
on 1369 Compaq
DS10L nodes

Why Linux?

- Free (speech & beer)
- Large developer community
- Kernel modules
  - No need to reboot during development
  - Supports partition model
- Supported on several platforms
- Familiarity with Linux
  - Ported Linux 2.0 to ASCI/Red

Results

- Cplant™ is now open source
- Large developer community is a wash
  - Most developers not focused on HPC and scaling issues
  - Extreme Linux helped
  - Extreme Linux isn't very extreme
- Modules
  - Big help in developing the networking stack
- Portals over any network device
  - Myrinet
  - RTS/CTS to skbufs
  - Portals over IP
  - Portals over IP in kernel
- Cplant™ runs on Alpha, x86, IA-64
- Linux changes too often to really be familiar

Other Observations

- Reliability
  - Linux hasn't been the cause of any machine interrupts
    - Still have other problems
  - Main selling point of Linux for the server market
- System software debugging tools are limited
- Application development environment more extensive
  - Compilers, debuggers, tools
- Lots of stuff we don't have to worry about
  - Device drivers: Ethernet, Serial
  - BIOS's
  - Hardware bugs
- Linux works OK for Cplant™ and commodity-based clusters

Technical Issues

- Predictability – avoid work unrelated to the computation
  - Linux on Alpha takes 1000 interrupts per second to keep time
  - Daemons: init, inetd, ipciod
  - Kernel threads: kswapd, kflushd, kupdate, kpiod

  Inappropriate resource management strategies
- VM system
  - Adverse impact on message passing
  - No physically contiguous memory
  - Must pin memory pages
  - Must maintain page tables on NIC
- Requires a filesystem
  - Not appropriate for diskless compute nodes

Social Issues

- Kernel development moves fast
  - Significant resources needed to keep up
- Distributions and development environments also change frequently
  - Tool vendors have trouble keeping up
- Linus changed out the VM system in the middle of the 2.4 kernels!
  - 2.4.9 – van Riel VM system
  - 2.4.10 – Arcangeli VM system
    - 150+ patches to the van Riel VM system
  - Linux fork?
- Server vs. multimedia desktop
  - Not HPC

Summary

- Linux works OK for Cplant™ and commodity clusters
  - CPU performance is acceptable for cluster bytes-to-FLOPS ratio
- Probable performance issues for platforms with a reasonable bytes-to-FLOPS ratio
- Community is a mixed blessing
  - Linux trades performance for everything

Future

- Take the good, leave the bad
- Leverage Linux hardware support - portability
  - BIOS
  - PCI chipsets
  - Processors
- Leverage application development environment
  - Compilers
  - Linkers
  - Binary file format
- Customize resource management strategies for HPC
  - Simple strategies have worked well